1 Model Description

Here, we describe the model we will use to model side effect function values. First, we note that we will use completely independent models for each side effect and treatment combination. While we did not have to adopt such an approach (for example, a patient's function might have a patient-specific and treatment-specific component), we do so for simplicity's sake.

In short, we assume that each patient has a 'true' function curve $g_i(t)$, and that the observed function values $g_i^*(t_j)$ are normally distributed about $g_i(t_j)$. The 'true' function curve will be parameterized by 3 latent parameters, each of which is modelled with a generalized linear model that depends on patient covariates. In the prior, the parameters for these 3 geenralized linear models are independent, though in the posterior, they will become dependent through the observed function values.

1.1 Parametric Form of Curve

Recalling the general function curve shapes from the previous section, we decide the 3 things we want to model of the curve $g_i(t)$ are:

1. The long term drop in function value
2. The short term drop in function value
3. The rate at which the function value recovers from the initial drop to the long term value.

Furthermore, we place the following restrictions on the curve:

1. The long term drop in function value is indeed a drop; that is, the long term function value is less than the pre-treatment function value
2. The short term drop is greater than the long term drop in function value.
3. 'True' function values at all times are between 0 and 1.

With these considerations in mind, we are ready to describe the parameterization of $g(t)$:

$$
\begin{aligned}
g(t; s, a, b, c) &= s(1 - a - b(1 - a)e^{-ct}) && (1)\\
\text{where} & && (2)\\
a &\in (0, 1) && (3)\\
b &\in (0, 1) && (4)\\
c &\in (0, \infty) && (5)\\
& && (6)
\end{aligned}
$$

and

- s is the pre-treatment function level
- a is the long term loss in function level, relative to the pre-treatment function level
- b is the short term loss in function level in excess of the long term loss, expressed as a portion of the long term function level
- c is the rate of the function level decays from the short term to long term function level

Note that the restrictions we place on the curve are enforced by the restrictions on a, b, c.

1.2 Model for Curve Parameters

Now that we defined how the 'true' function level curve is parameterized, we will define the models for each of the parameters. As the parameters' ranges are constrained, we cannot use a standard linear model for the parameters. However, generalized linear models will suit us perfectly. Furthermore, as a and b have the same range, the models for a and b will be analogous to each other.

1.2.1 Generalized Linear Model

Standard linear regression models the observed variable Y as coming from a Normal distribution such that $E(Y)$ is a linear function of the covariate vector X. That is, $Y \sim N(BX, \sigma)$, where σ is the standard deviation of Y, and more generally, a parameter that either directly or indirectly specifies the spread of Y. Generalized linear regression generalizes linear regression in 2 ways: 1. The observed variable Y comes from a distribution with a density $p(Y)$ that is not necessarily that of the normal distribution, and 2. $E(Y)$ is no longer necessarily a linear function of covariates, but the result of a linear function of covariates that is subsequently sent through a link function. That is, $E(Y) = f(BX)$, where f is the link function. Specification of f allows one to control the manner in which the mean response $E(Y)$ depends on covariates. In particular, it allows one to restrict the range of $E(Y)$.

A Generalized linear model relating the observed variable Y to covariate vector X thus has 2 components:

- a probability distribution $p(Y)$ parameterized, its mean $\mu = E(Y)$ and a parameter ϕ controlling the variance of Y.
- a link function f such that $\mu = f(BX + z)$

Note that we explicitly allow for a bias term z. For example, linear regression can be thought of as a generalized linear model where $p(Y) = p_{\text{normal}}(y; \mu = \mu, \sigma = \phi)$, and the link function f being the identity function.

1.2.2 Model for a

As a must reside in the unit interval, if we are to model a using a GLM, $p(a : \mu, \phi)$ must have zero support outside the unit interval. The Beta distribution is a good candidate for $p(a)$.

1.2.3 Beta Distribution

This is a continuous distribution with support in (0,1). There are various parameterizations of the Beta distribution. The most common one, for a distribution $Beta(\alpha, \beta)$ is as such:

$$p_{Beta}(y; \alpha, \beta) \propto y^{\alpha-1}(1-y)^{\beta-1} \tag{7}$$

with $E(X) = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. However, for GLM, we require $p(Y)$ to be parameterized by $\mu = E(Y)$ and some dispersion parameter ϕ. Fortunately, some properties of the Beta distribution allow for such a parameterization. Firstly, if we let

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{8}$$

$$\phi = \frac{1}{\alpha + \beta + 1} \tag{9}$$

we see that a $Beta(\alpha, \beta)$ random variable has mean μ and variance $\mu(1-\mu)\phi$. We can then solve the above 2 equations for μ and ϕ in terms of α and β. Then, we can obtain an alternate parameterization for a Beta random variable $p(y; \mu, \phi) =$

1.3 GLM for a

Now, we can describe the model for a_i concisely:

$$a_i \sim Beta(\mu_i^a, \phi^a) \tag{10}$$

$$\mu_i^a = f^a(B^a X_i + z^a) \tag{11}$$

$$f^a(x) = \frac{1}{1 + e^{-x}} \tag{12}$$

f^a is the link function for the GLM, which we have chosen to be the logistic function, as μ_i^a must be between 0 and 1. We have not described what z^a will be yet, but will do so shortly. b is defined analogously to a. We have not defined what distribution ϕ^a follows. For now, the important thing is that ϕ^a is shared between patients, and is between 0 and 1.

1.4 GLM for c

Keeping in mind that $c > 0$, we will let c come from a gamma distribution. The gamma distribution can be parameterized by its mean μ and a shape parameter k that controls the variance of the distribution. We would like c to come from a unimodal distribution, which will be the case if $k > 1$. We model the inverse of k instead, letting $\phi^c = \frac{1}{k}$ in the traditional gamma distribution parameterization, keeping in mind that k is between 0 and 1, and is again, shared between patients.

Thus, the model for c_i is as such:

$$\begin{aligned} c_i &\sim Gamma(\mu_i^c, \phi^c) && (13)\\ \mu_i^c &\sim f^c(B^c X_i + z^c) && (14)\\ f^c(x) &= e^x && (15) \end{aligned}$$

Once again, we defer specification of the priors for the parameters z^c, B^c and ϕ^c until later.

1.5 Data Normalization and choosing intercepts z^a, z^b, z^c

The way we normalize the data and choose the intercept parameters is guided by our desire in the model that the the 'average' patient should, under our model, expect to receive the average curve as their 'true' curve, which is a curve for which a, b, c are equal to the average value of those parameters in our dataset, which we will denote as $\mu_{pop}^a, \mu_{pop}^b, \mu_{pop}^c$, respectively. We would like, for the patient with the average covariate vector X_{pop}, $E(a) = \mu_{pop}^a$, and likewise for b and c. We can accomplish by doing the following:

1. normalize each covariate to have 0 mean and standard deviation 1 across the dataset
2. Set $z^a = g^{a-1}(\mu_{pop}^a)$

Looking at the equation for μ_i^a, one sees that this way, the average patient has a covariate vector equal to the 0 vector, and receives a value of $\mu_i^a = \mu_{pop}^a$.

1.6 Extracting mean population parameters $\mu_{pop}^a, \mu_{pop}^b, \mu_{pop}^c$

To calculate these mean parameters, we need to know their values for each patient. But we don't actually observe patient curve parameters a, b, c, and so we will obtain them through least squares curve fitting, ensuring that the curve parameters obey their constraints. More specifically, for each patient i, we want to:

$$\begin{aligned} \text{minimize} \sum_{j=1}^{m} (g_i^*(t_j) &- g_i(t_j; s_i, a_i, b_i, c_i))^2 \\ \text{subject to} & \\ a_i &\in (0, 1) \\ b_i &\in (0, 1) \\ c_i &\in (0, \infty) \end{aligned}$$

For curves where the function value does not drop (or even increases) after treatment, a will be quite small or even 0, and the fitted curve may be flat when the actual curve rises. However, we view this as unavoidable due to the restrictions of the kinds of curve we allow in our model.

1.7 Model for observed data

We have described how a patient's true curve $g(t; s, a, b, c$depends on 3 latent parameters a, b, c and observed pre-treatment value s, and how those 3 parameters depend on patient covariates and the parameters of the model B_a, B_b, B_c. Now, it remains to specify how the observed data $g^*(t_j; s, a, b, c)$ depends on $g(t_j; a, b, c)$. We will take a simplistic model. We will assume that the observed function value at time t_j is normally distributed about the true function value. That is,

$$g_i^*(t_j; s_i, a_i, b_i, c_i) \sim N(g_i(t_j; s_i, a_i, b_i, c_i, \sigma^{noise}) \quad (16)$$

for all patients i and at all measurement times $t_1, \ldots, t_j$.

Thus, we are assuming that the observed function values for a patient are conditionally independent of each other given the patient's true curve parameters. In reality perhaps those observed values might be correlated, but we opt for simplicity whenever possible.

1.8 Priors for parameters

Thus far, our model contains 7 parameters: $\theta = \{B_a, B_b, B_c, \phi^a, \phi^b, \phi^c, \phi^{noise}\}$. As we adopt a Bayesian framework, we must give prior distributions for each of those parameters $P(\theta; \alpha)$ where α is a set of hyperparamters.

1.8.1 Desired prior predictive distribution for a, b, c

To describe what properties we want of $P(\theta; \alpha)$, it is more useful to describe, before observing any data, what we want the distribution over $g(t; s, a, b, c)$ to look like. As $g(t; s, a, b, c)$ is fully described by a, b, c, then what we want to do is, for an test patient X, describe the prior predictive distributions $P(a; \alpha, X), P(b; \alpha, X), P(c; \alpha, X)$. X denotes the test sample, not the data that we have yet to observe. These distributions will of course depend on X. We want $P(a; \alpha, X)$ to be:

1. unimodal, for reasonably values of X
2. roughly centered around μ^a_{pop}
3. to have larger variance the further X is from the average covariate vector. (since we are normalizing the data to be mean 0, this means the larger the magnitude of X, the larger the variance of $P(a; \alpha, X)$.

These same desirables apply to the prior predictive distributions of b and c.

1.8.2 Specifying the Priors for Parameters

We will let the parameters B_a, B_b, B_c each follow normal distributions in the prior. These normal distributions will be be mean zero, and have a diagonal covariance matrix equal to some scalar multiple of the identity matrix. That is, we will let:

$$
\begin{aligned}
B_a &\sim N(0, c^a I) && (17)\\
B_b &\sim N(0, c^b I) && (18)\\
B_c &\sim N(0, c^c I) && (19)\\
& && (20)
\end{aligned}
$$

where c^a, c^b, c^c are hyperparameters, as they describe prior distributions of model parameters. We will also let the disperson parameters $\phi^a, \phi^b, \phi^c, \phi^{noise}$ follow exponential distributions truncated at 1. That is, we will let:

$$
\begin{aligned}
\phi^a &\sim \text{truncated\_exp}(\lambda^a, 1) && (21)\\
\phi^b &\sim \text{truncated\_exp}(\lambda^b, 1) && (22)\\
\phi^c &\sim \text{truncated\_exp}(\lambda^c, 1) && (23)\\
\phi^{noise} &\sim \text{truncated\_exp}(\lambda^{noise}, 1) && (24)\\
& && (25)
\end{aligned}
$$

The reason why we choose the prior for these parameters is that we want to encourage them to be small so that the distributions for a, b, c will have relatively small variance in the prior. The reason for this is that in general, distributions with high variances with finite support will not be unimodal, which is a situation we want to avoid.

Now, we analyze the influence of the hyperparameters c^a and λ^a on the prior predictive distribution $P(X|\alpha)$. We will do so in 2 steps: First study how $P(X|c^a)$ depends on X and c^a.

1.9 Distribution of $P(\mu^a|c^a, \tilde{X})$